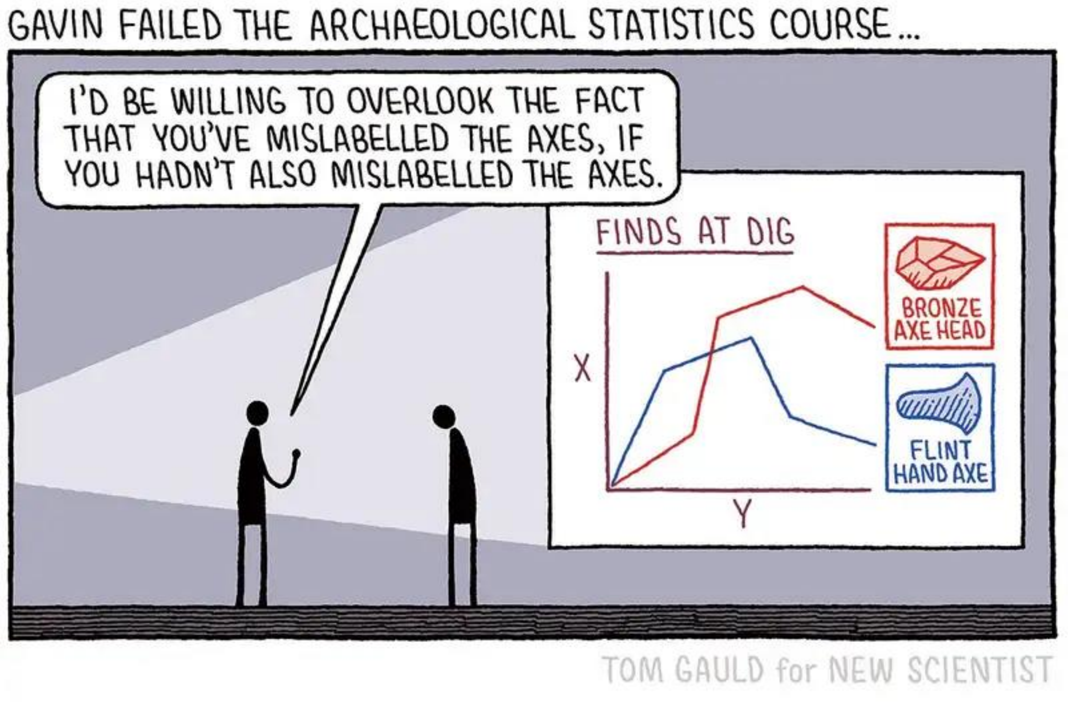
# Quantifying Culture

(Class 3)

**Introducing inferential statistics**


GAVIN FAILED THE ARCHAEOLOGICAL STATISTICS COURSE...
I'D BE WILLING TO OVERLOOK THE FACT THAT YOU'VE MISLABELLED THE AXES, IF YOU HADN'T ALSO MISLABELLED THE AXES.
FINDS AT DIG
X
Y
BRONZE AXE HEAD
FLINT HAND AXE
TOM GAULD for NEW SCIENTIST

UNIVERSITÀ DI TORINO

Alessio Palmisano
University of Turin

Tech 4 Culture

# Outline

- Linear regression
- Chi-square test
- Principal component analysis (PCA)

# Linear regression

- Regression

A statistical method used to model and analyse relationships between dependent and independent variables (covariates).

- Correlation

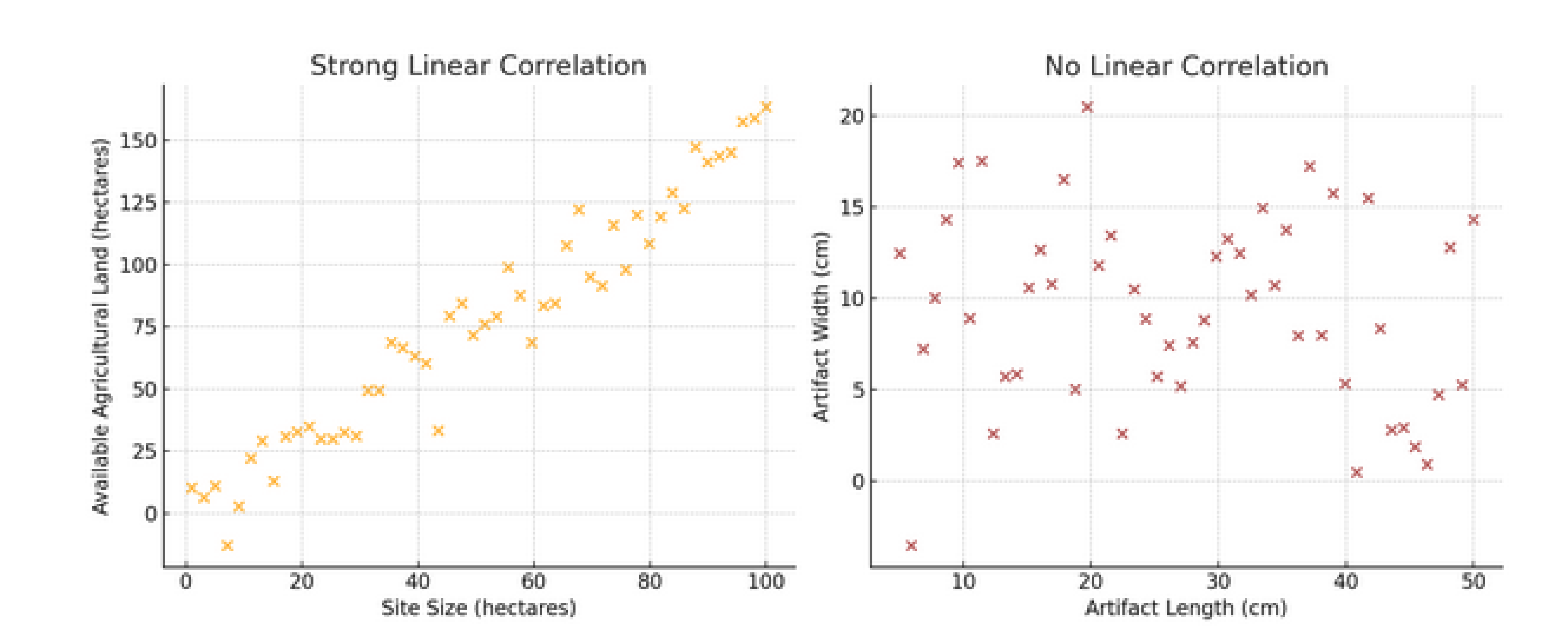
A statistical measure describing the relationship between two variables—how well they move together.

**Correlation** tells you how strongly two variables are related.

*VS*

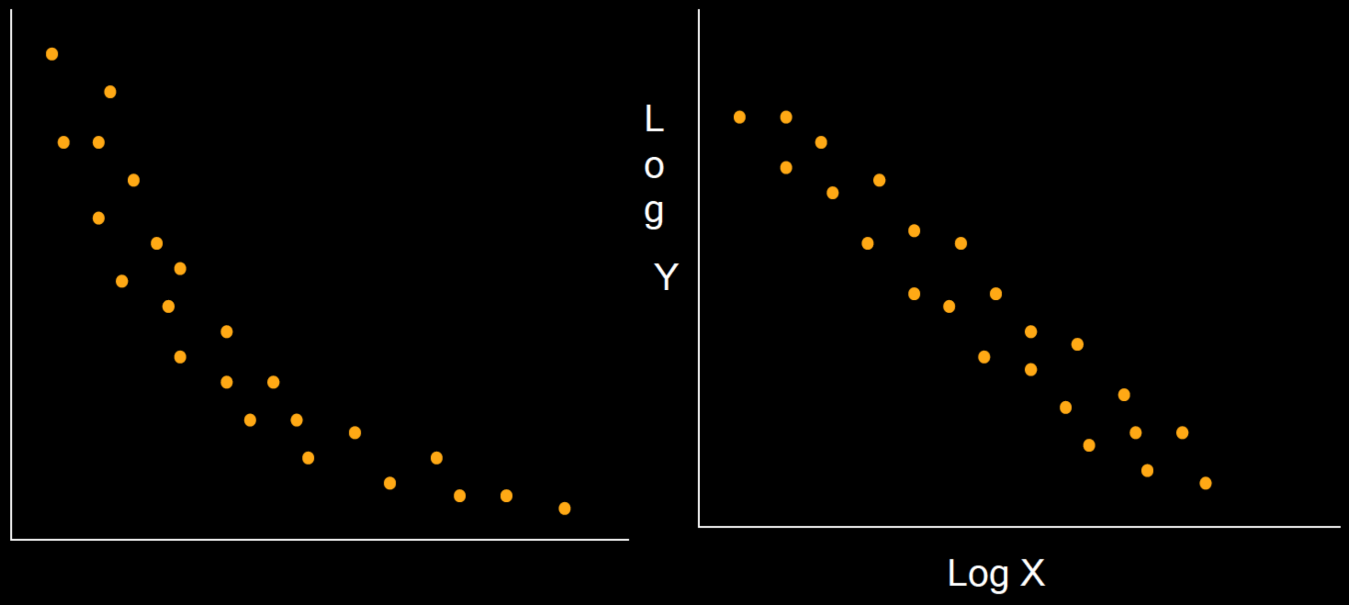
**Regression** helps predict one variable based on another.

# Correlation



# Correlation

In some cases you need to linearise a relationship through the **Log-transformation** of the original variables’ values

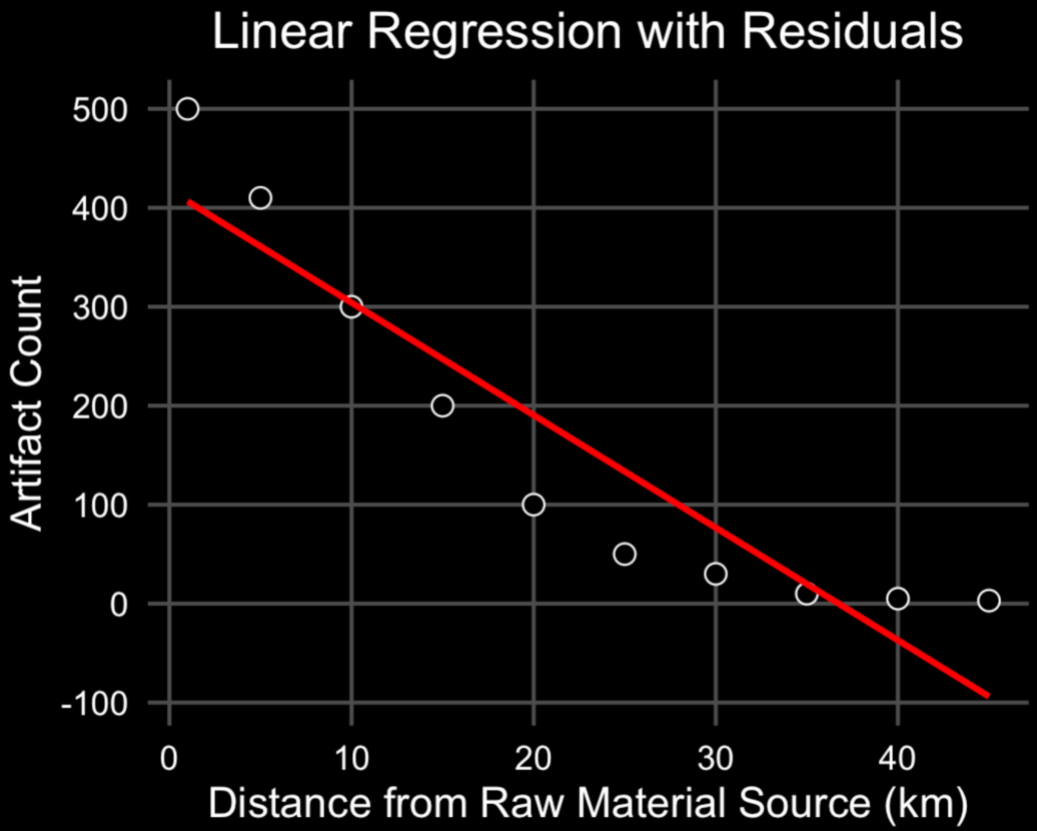


# Linear regression

$$Y = a + bx$$

$a$ = The **intercept** is the point where the regression line crosses the **Y-axis** when the **independent** variable is zero.

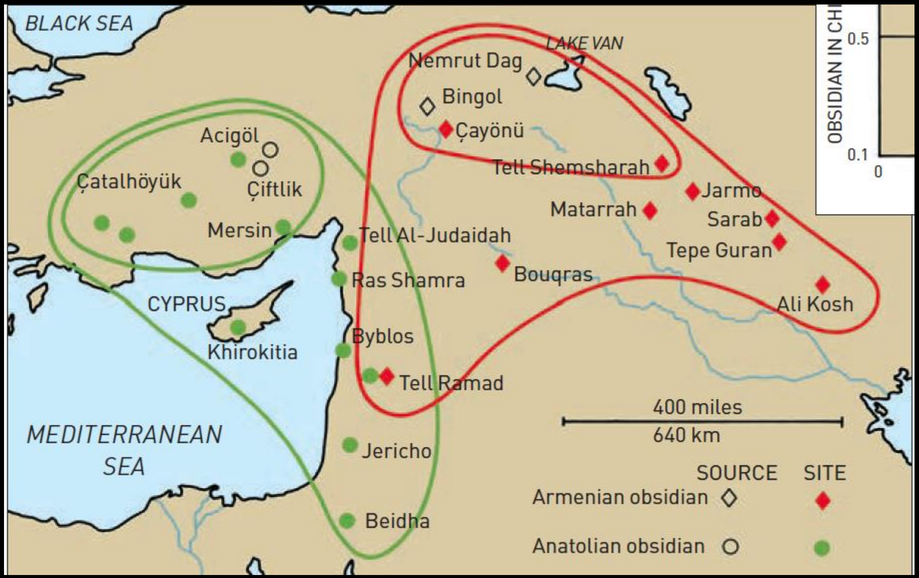
$b$ = *the slope* represents the change in *y* for each **one-unit increase** in *x*. It indicates how strongly YYY depends on XXX.

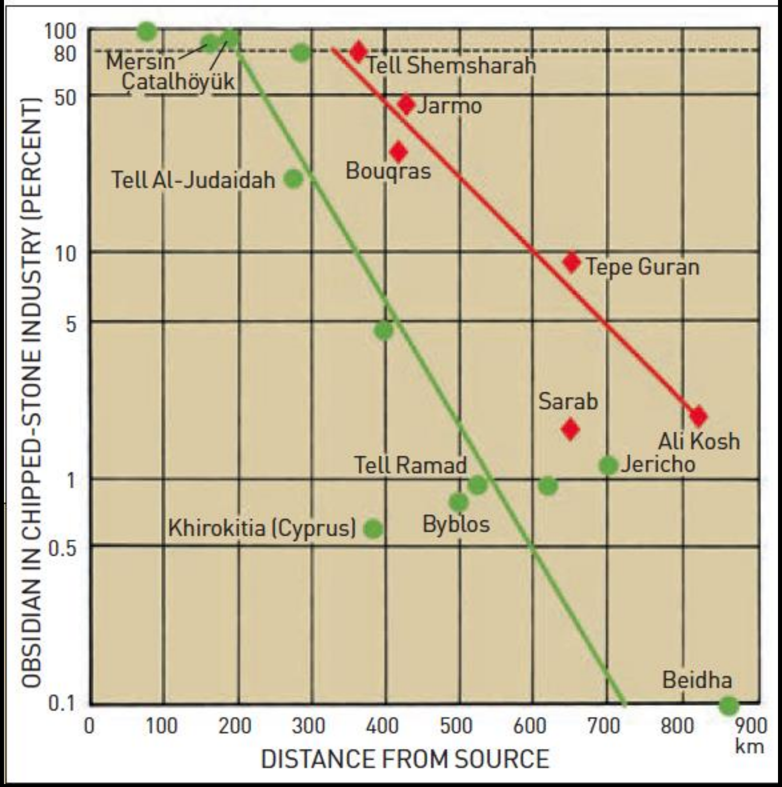


$Y$ = 417 + -11.37 * x (distance from material source)

# Linear regression – case study

$$Y = a + bx$$





Modelling the decline obsidian artifacts with increasing distance from source area.

# Correlation

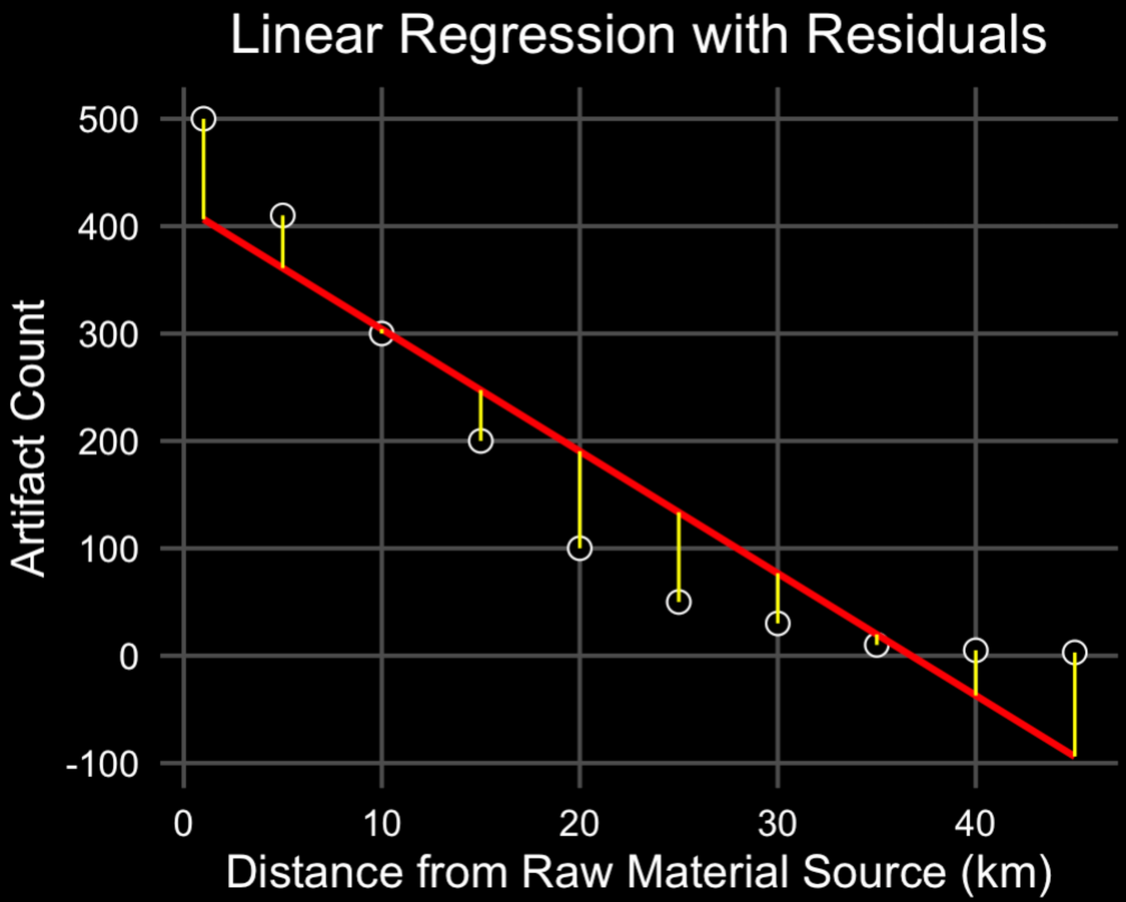
The numbers of artifact are negatively correlated with the distance from raw material source.

- Pearson *r* correlation = 0.92

- Coefficient of determination $r^2 = 0.84$

It indicates that 84% of the variation in the artifact count can be explained by the distance from material source.

- Residuals



# Coding time

- Tutorial pp. 1 - 16

# Chi-square test

It is a statistical method used to determine whether there is a significant **association** between **two categorical** variables.

## One or two-sample versions

It helps answer questions like:

- Are two categorical variables independent of each other?
- Does the observed data match the expected distribution?

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

**O** = Observed frequency (actual counts in your sample)
**E** = Expected frequency (counts we would expect under the assumption of independence)

# Chi-square test (one sample test)

**Null hypothesis ($H_o$):** no association between variables.

**Alternative hypothesis ($H_1$):** association between variables.

For each category

Calculate expected frequencies:

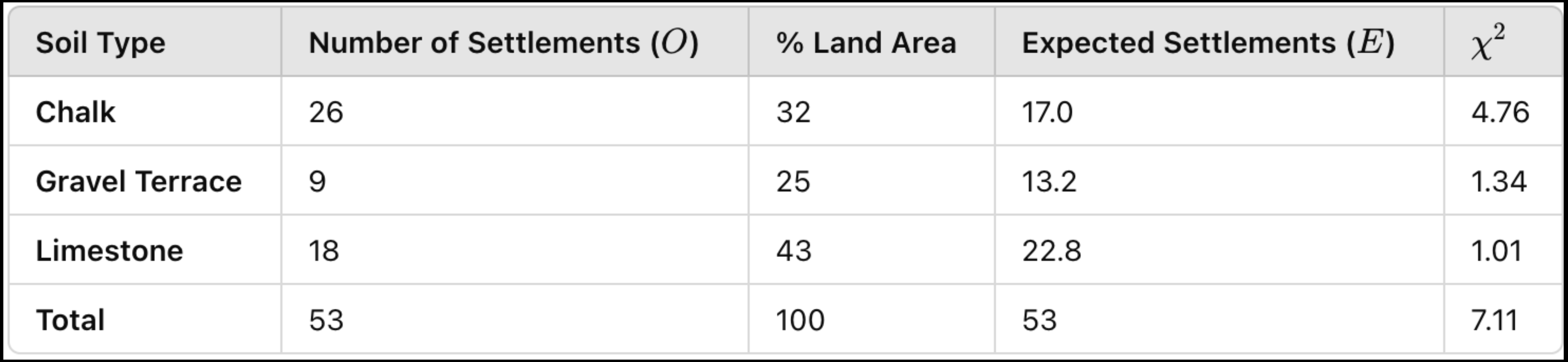
$$E = \frac{(\text{Total Settlements} \times \%\ \text{Land Area})}{100}$$

Calculate the chi-square:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Degrees of freedom (v) = (no. of rows – 1)(no. of columns – 1)

Case study : Iron Age inhumation cemetery in Germany (Shennan 1997, 65-70).
Testing the relationship between individual’s sex and the side on which that individual is lying in the grave.



| Soil Type | Number of Settlements ($O$) | % Land Area | Expected Settlements ($E$) | $\chi^2$ |
|---|---|---|---|---|
| **Chalk** | 26 | 32 | 17.0 | 4.76 |
| **Gravel Terrace** | 9 | 25 | 13.2 | 1.34 |
| **Limestone** | 18 | 43 | 22.8 | 1.01 |
| **Total** | 53 | 100 | 53 | 7.11 |

Shennan, S. (1997). *Quantifying archaeology*. University of Iowa Press.

# Chi-square test (one sample test)

**Null hypothesis ($H_o$):** no association between variables.

**Alternative hypothesis ($H_1$):** association between variables.

For each category

Calculate expected frequencies:

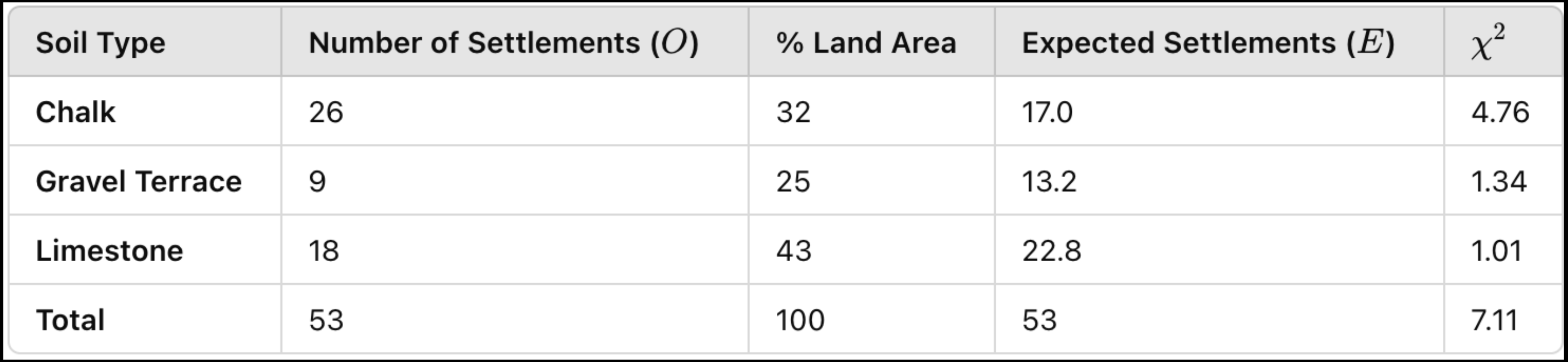
$$E = \frac{(\text{Total Settlements} \times \%\ \text{Land Area})}{100}$$

Calculate the chi-square:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Degrees of freedom (v) = (no. of rows – 1)(no. of columns – 1)

Case study : Iron Age inhumation cemetery in Germany (Shennan 1997, 65-70).
Testing the relationship between individual’s sex and the side on which that individual is lying in the grave.



| Soil Type | Number of Settlements ($O$) | % Land Area | Expected Settlements ($E$) | $\chi^2$ |
|---|---|---|---|---|
| **Chalk** | 26 | 32 | 17.0 | 4.76 |
| **Gravel Terrace** | 9 | 25 | 13.2 | 1.34 |
| **Limestone** | 18 | 43 | 22.8 | 1.01 |
| **Total** | 53 | 100 | 53 | 7.11 |

In this example, at the significance level of =0 .05, the test statistic (7.11) is greater than the required value (5.99) and hence the pattern is a significant one.

# Chi-square test

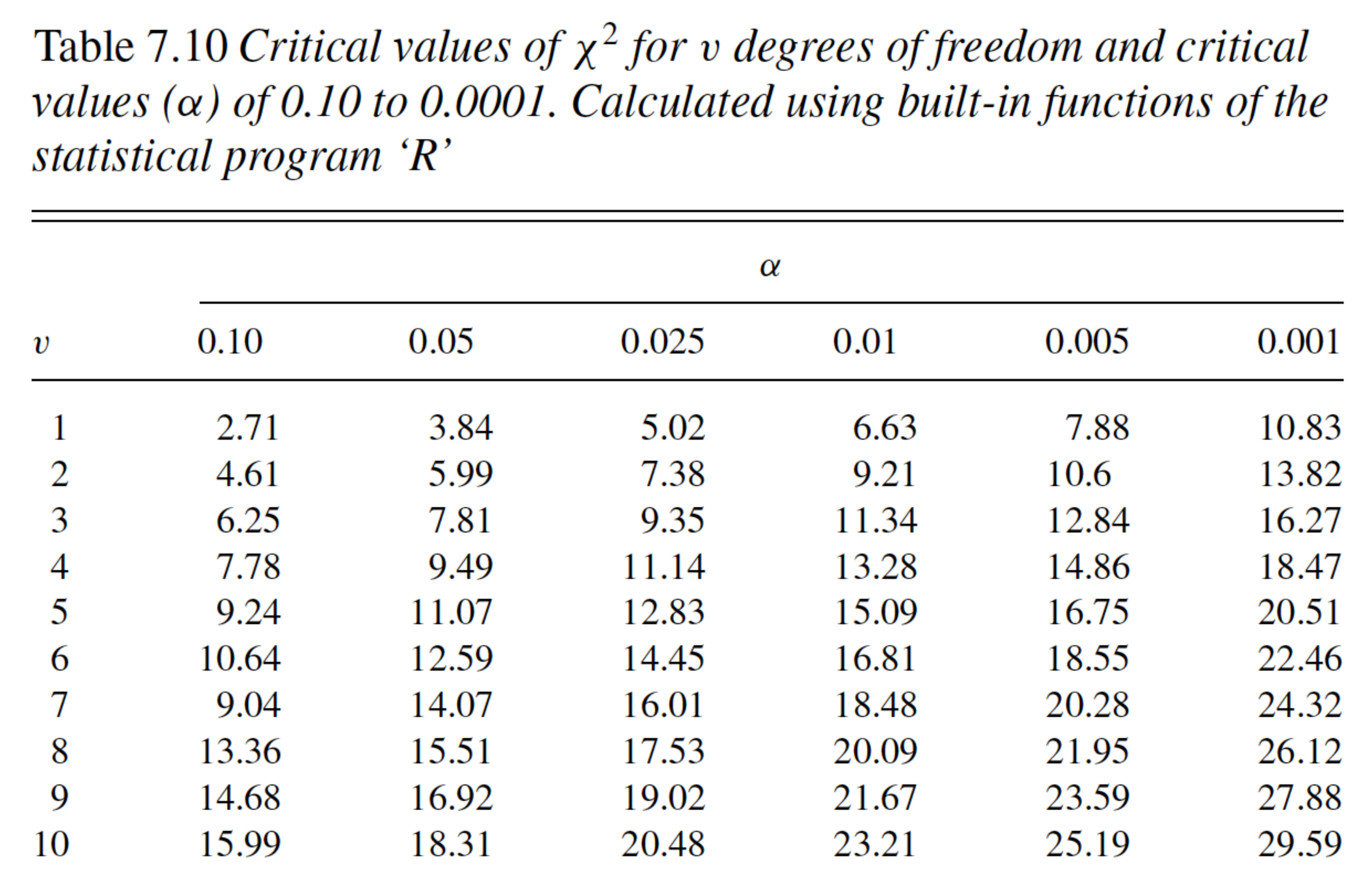


Table 7.10 *Critical values of $\chi^2$ for v degrees of freedom and critical values (α) of 0.10 to 0.0001. Calculated using built-in functions of the statistical program 'R'*

|  | $\alpha$ |  |  |  |  |  |
|---|---|---|---|---|---|---|
| $v$ | 0.10 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 |
| 1 | 2.71 | 3.84 | 5.02 | 6.63 | 7.88 | 10.83 |
| 2 | 4.61 | 5.99 | 7.38 | 9.21 | 10.6 | 13.82 |
| 3 | 6.25 | 7.81 | 9.35 | 11.34 | 12.84 | 16.27 |
| 4 | 7.78 | 9.49 | 11.14 | 13.28 | 14.86 | 18.47 |
| 5 | 9.24 | 11.07 | 12.83 | 15.09 | 16.75 | 20.51 |
| 6 | 10.64 | 12.59 | 14.45 | 16.81 | 18.55 | 22.46 |
| 7 | 9.04 | 14.07 | 16.01 | 18.48 | 20.28 | 24.32 |
| 8 | 13.36 | 15.51 | 17.53 | 20.09 | 21.95 | 26.12 |
| 9 | 14.68 | 16.92 | 19.02 | 21.67 | 23.59 | 27.88 |
| 10 | 15.99 | 18.31 | 20.48 | 23.21 | 25.19 | 29.59 |

# Chi-square test (two sample test)

**Null hypothesis ($H_o$):** no association between variables.

**Alternative hypothesis ($H_1$):** association between variables.

For each category

Calculate expected frequencies:

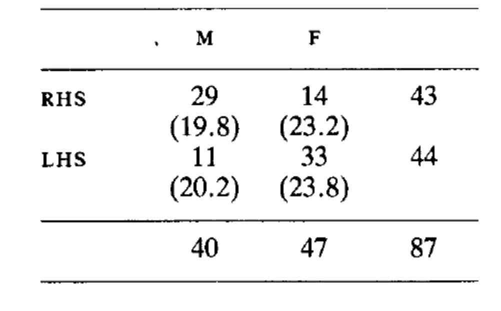
$$E = \frac{(\text{Row Total} \times \text{Column Total})}{\text{Grand Total}}$$

Calculate the chi-square:

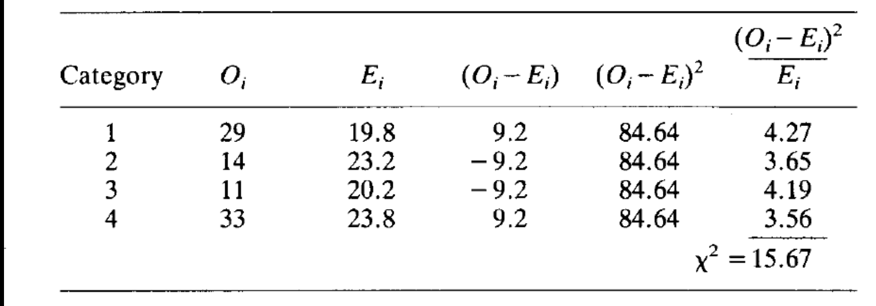
$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Degrees of freedom (v) = (no. of rows – 1)(no. of columns – 1)

Case study : Iron Age inhumation cemetery in Germany (Shennan 1997, 70-74).
Testing the relationship between individual's sex and the side on which that individual is lying in the grave.



| | M | F | |
|---|---|---|---|
| RHS | 29<br>(19.8) | 14<br>(23.2) | 43 |
| LHS | 11<br>(20.2) | 33<br>(23.8) | 44 |
| | 40 | 47 | 87 |



| Category | $O_i$ | $E_i$ | $(O_i - E_i)$ | $(O_i - E_i)^2$ | $\frac{(O_i - E_i)^2}{E_i}$ |
|---|---|---|---|---|---|
| 1 | 29 | 19.8 | 9.2 | 84.64 | 4.27 |
| 2 | 14 | 23.2 | −9.2 | 84.64 | 3.65 |
| 3 | 11 | 20.2 | −9.2 | 84.64 | 4.19 |
| 4 | 33 | 23.8 | 9.2 | 84.64 | 3.56 |
| | | | | | $\chi^2 = 15.67$ |

Shennan, S. (1997). *Quantifying archaeology*. University of Iowa Press.

# Chi-square test (two sample test)

**Null hypothesis ($H_o$):** no association between variables.

**Alternative hypothesis ($H_1$):** association between variables.

For each category

Calculate expected frequencies:

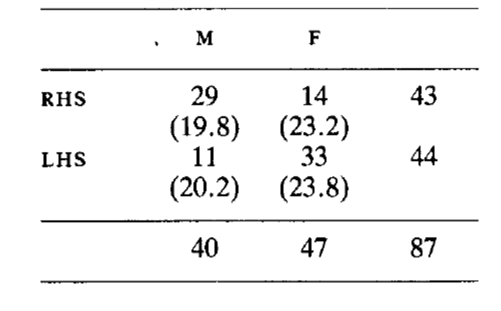
$$E = \frac{(\text{Row Total} \times \text{Column Total})}{\text{Grand Total}}$$

Calculate the chi-square:

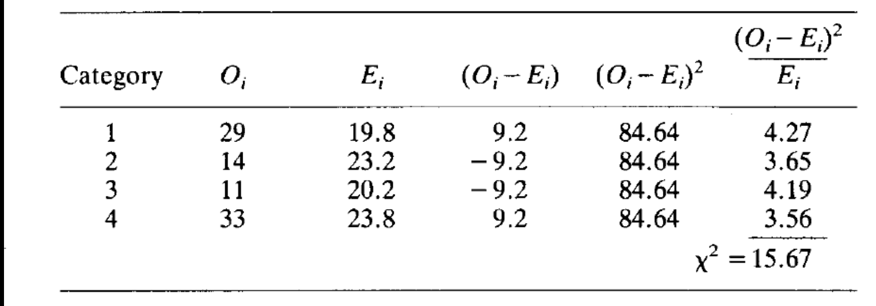
$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Degrees of freedom (v) = (no. of rows – 1)(no. of columns – 1)

Case study : Iron Age inhumation cemetery in Germany (Shennan 1997, 70-74).
Testing the relationship between individual's sex and the side on which that individual is lying in the grave.



|  | M | F |  |
|---|---|---|---|
| RHS | 29<br>(19.8) | 14<br>(23.2) | 43 |
| LHS | 11<br>(20.2) | 33<br>(23.8) | 44 |
|  | 40 | 47 | 87 |



| Category | $O_i$ | $E_i$ | $(O_i - E_i)$ | $(O_i - E_i)^2$ | $\frac{(O_i - E_i)^2}{E_i}$ |
|---|---|---|---|---|---|
| 1 | 29 | 19.8 | 9.2 | 84.64 | 4.27 |
| 2 | 14 | 23.2 | −9.2 | 84.64 | 3.65 |
| 3 | 11 | 20.2 | −9.2 | 84.64 | 4.19 |
| 4 | 33 | 23.8 | 9.2 | 84.64 | 3.56 |
|  |  |  |  |  | $\chi^2 = 15.67$ |

In this example, at the significance level of =0 .05, the test statistic (15.67) is greater than the required value (3.84) and hence the pattern is a significant one.

# Coding time

- Tutorial pp. 16 - 19

# Principal Component Analysis (PCA)

It is a mathematical technique used to **reduce the dimensionality** of a dataset while **preserving** as much **variability (information)** as possible.

Transforming the data into a new set of variables called **principal components (PCs)**, which are ordered by the amount of variance they explain in the data.

# Principal Component Analysis (PCA)

Why?

1. **Dimensionality Reduction** – Reducing the number of variables while maintaining the important patterns in the data.

2. **Data Visualization** – Making it easier to visualize high-dimensional data in 2D or 3D.

3. **Noise Reduction** – Removing less significant components that might represent noise.

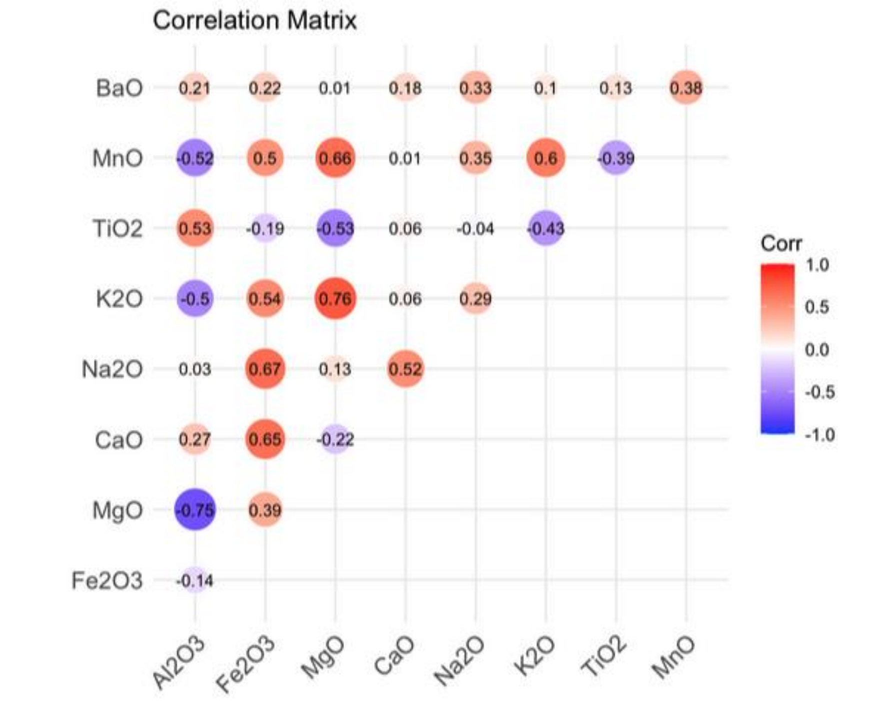
4. **Feature Extraction** – Creating new features that are linear combinations of the original ones.

# Principal Component Analysis (PCA)

## Steps

1. **Standardisation of data (optional)** – to standardize the dataset so that all variables contribute equally.

2. **Compute a covariance matrix** – to show how different variables in the dataset are related to each other.



| | $Al_2O_3$ | $Fe_2O_3$ | $MgO$ | $CaO$ | $Na_2O$ | $K_2O$ | $TiO_2$ | $MnO$ |
|---|---|---|---|---|---|---|---|---|
| $BaO$ | 0.21 | 0.22 | 0.01 | 0.18 | 0.33 | 0.1 | 0.13 | 0.38 |
| $MnO$ | -0.52 | 0.5 | 0.66 | 0.01 | 0.35 | 0.6 | -0.39 | |
| $TiO_2$ | 0.53 | -0.19 | -0.53 | 0.06 | -0.04 | -0.43 | | |
| $K_2O$ | -0.5 | 0.54 | 0.76 | 0.06 | 0.29 | | | |
| $Na_2O$ | 0.03 | 0.67 | 0.13 | 0.52 | | | | |
| $CaO$ | 0.27 | 0.65 | -0.22 | | | | | |
| $MgO$ | -0.75 | 0.39 | | | | | | |
| $Fe_2O_3$ | -0.14 | | | | | | | |

# Principal Component Analysis (PCA)

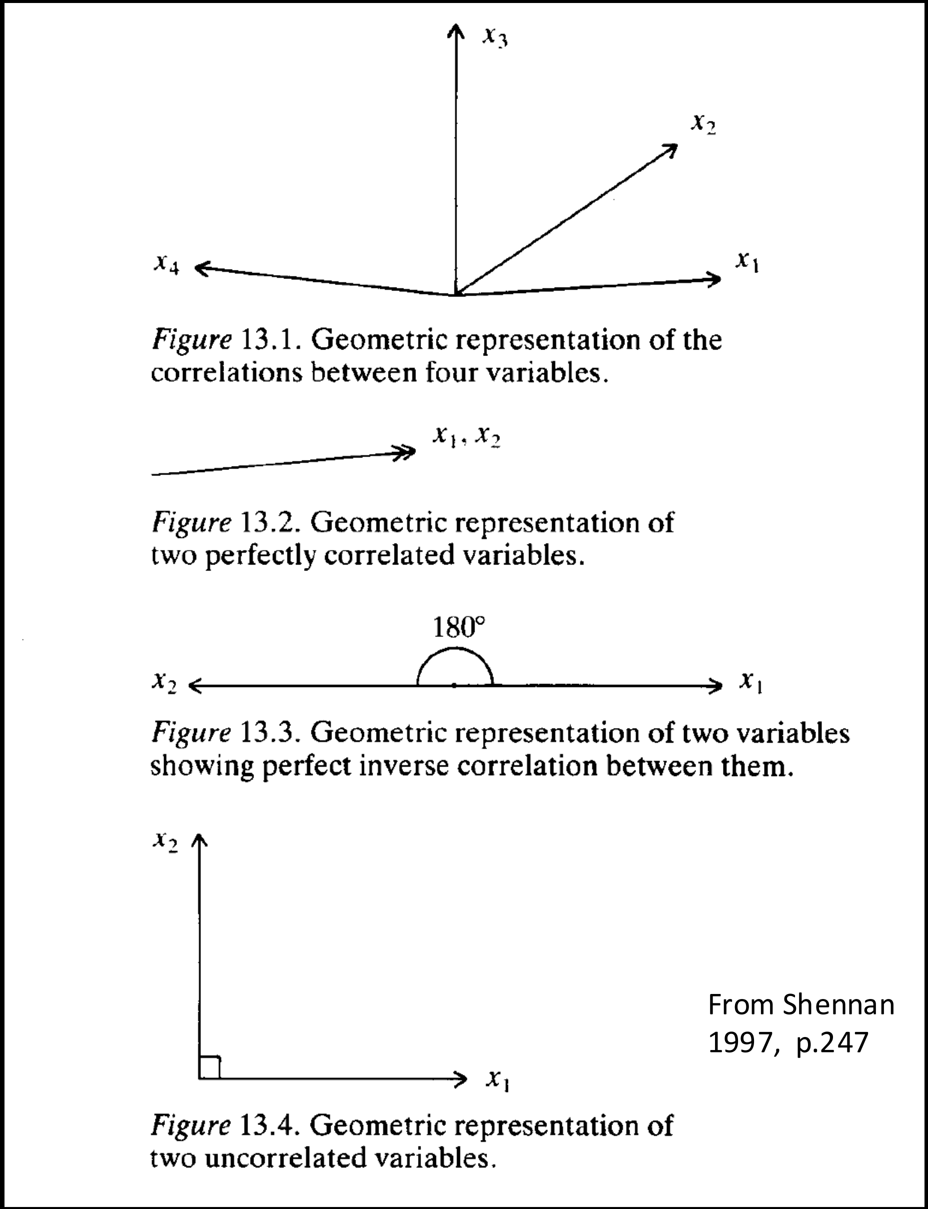


*Figure* 13.1. Geometric representation of the correlations between four variables.

*Figure* 13.2. Geometric representation of two perfectly correlated variables.

*Figure* 13.3. Geometric representation of two variables showing perfect inverse correlation between them.

From Shennan 1997, p.247

*Figure* 13.4. Geometric representation of two uncorrelated variables.

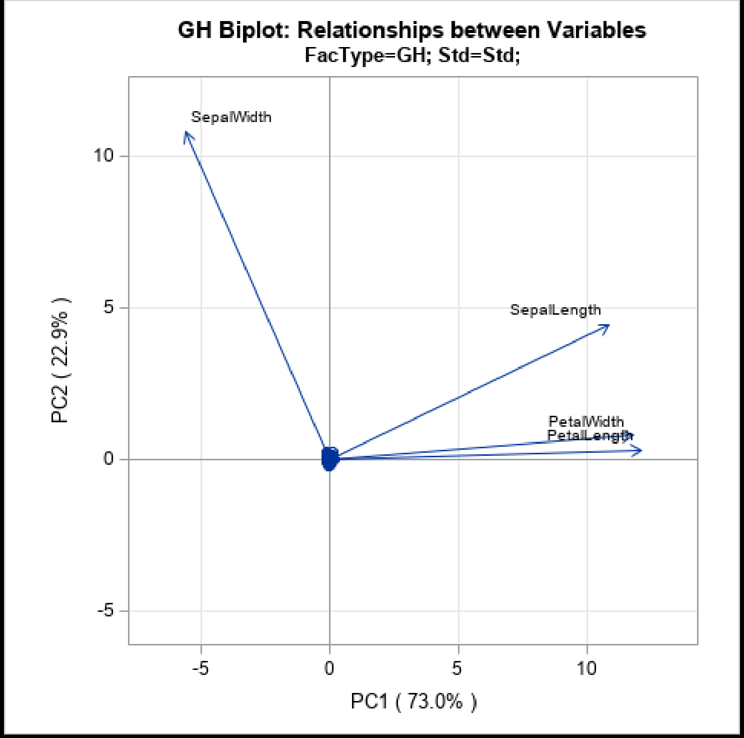
Shennan, S. (1997). *Quantifying archaeology*. University of Iowa Press.

# Principal Component Analysis (PCA)

## Steps

### 3. Compute the *loadings* of each variable

- They represent how much original variable **contributes** to the principal components.
- The loadings are shown **as vectors (arrows)** on a plot.
- **The length** of the arrow indicates the **magnitude** of the contribution of that variable to the principal component.
- The **direction** of the arrow indicates the variable's **relationship** with the principal component.

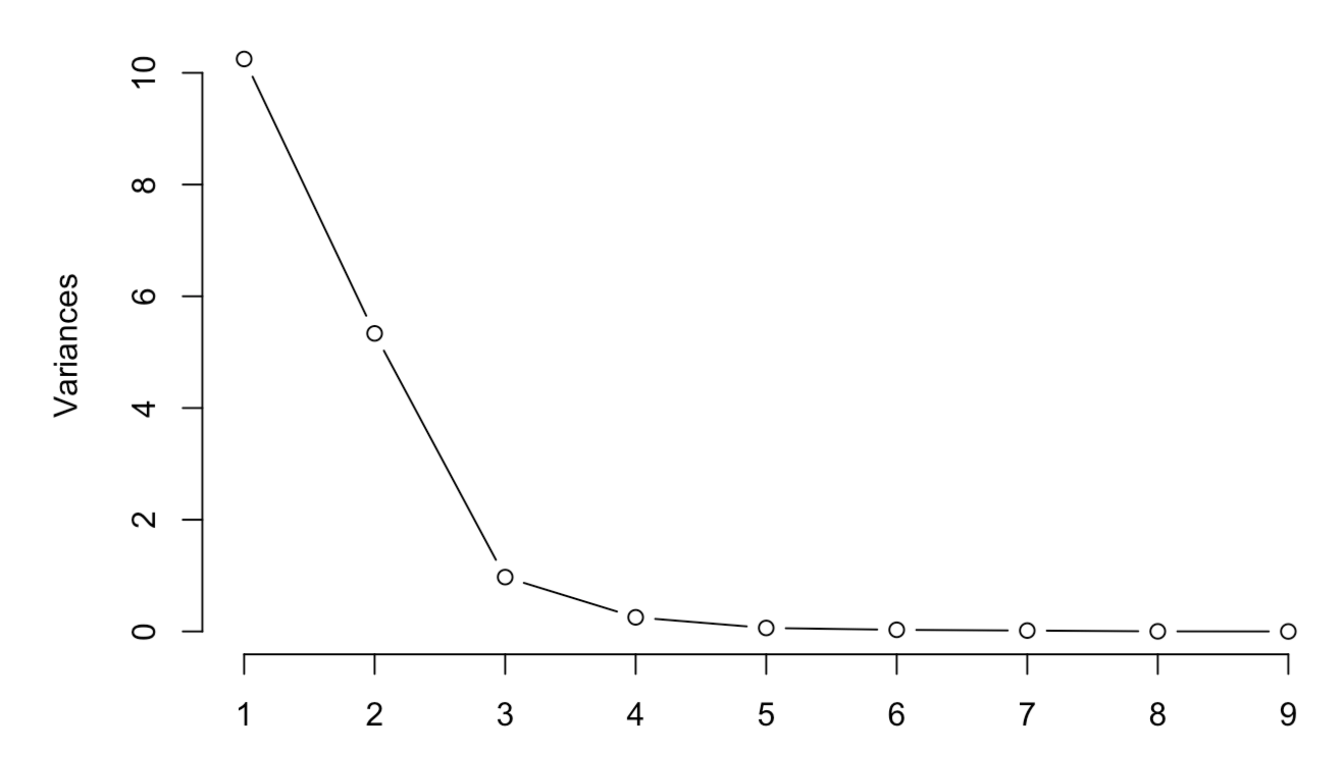


# Principal Component Analysis (PCA)

## Steps

### 4. Select the Top Principal Components

- Choosing the top principal components that explain most of the variance (often using a **threshold like 90%** of total variance).
- Using a scree plot to identify the optimal number pf components.

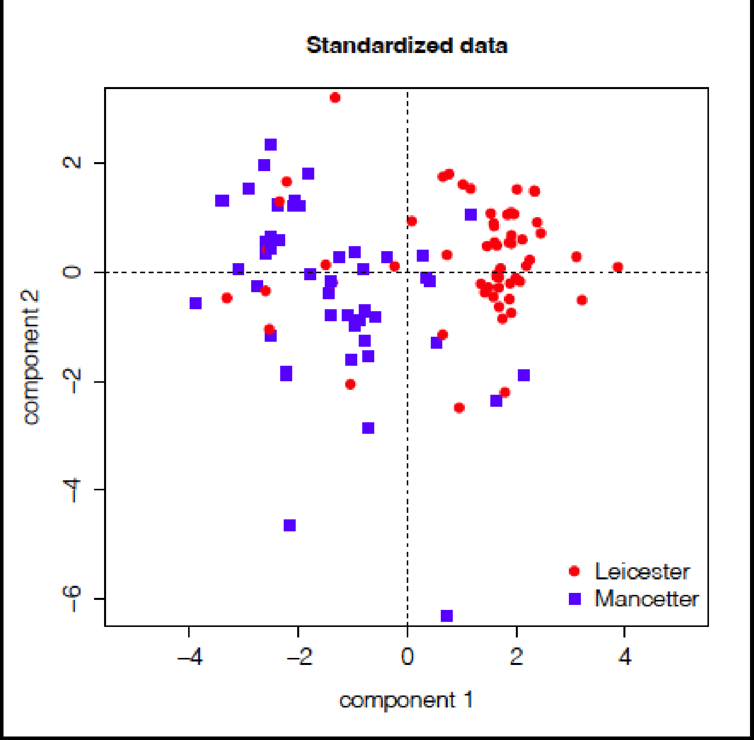


# Principal Component Analysis (PCA)

## Steps

### 5. Plot the PCA scatterplot

- The points represent the PCA scores of observed data according to their correlations with the principal components (PCs)
- Clusters, trends, or separations between different classes can become more apparent.



105 specimens of Romano-British glass waste from Leicester and Manchester (Baxter 2015, p.107)

**Hypothesis:** The glass from the two sites are chemically distinct.

Baxter, M., 2015. *Notes on Quantitative Archaeology and R.*

# Coding time

- Tutorial pp. 19 - 27